VALUE IN HEALTH 15 (2012) 167–174

Available online at www.sciencedirect.com
SciVerse ScienceDirect
journal homepage: www.elsevier.com/locate/jval

Extension of Kaplan-Meier Methods in Observational Studies with Time-Varying Treatment

Stanley Xu, PhD[1,2,*], *Susan Shetterly, MS*[1], *David Powers, MS*[1], *Marsha A. Raebel, PharmD*[1,2], *Thomas T. Tsai, MD, MSc*[1,2,3], *P. Michael Ho, MD, PhD*[1,2,3], *David Magid, MD, MPH*[1,2]

[1]*The Institute for Health Research, Kaiser Permanente Colorado, Denver, CO, USA;* [2]*University of Colorado, Denver, CO, USA;* [3]*Denver VA Medical Center, Denver, CO, USA*

ABSTRACT

Objectives: Inverse probability of treatment weighted Kaplan-Meier estimates have been developed to compare two treatments in the presence of confounders in observational studies. Recently, stabilized weights were developed to reduce the influence of extreme inverse probability of treatment-weighted weights in estimating treatment effects. The objective of this research was to use adjusted Kaplan-Meier estimates and modified log-rank and Wilcoxon tests to examine the effect of a treatment that varies over time in an observational study. **Methods:** We proposed stabilized weight adjusted Kaplan-Meier estimates and modified log-rank and Wilcoxon tests when the treatment was time-varying over the follow-up period. We applied these new methods in examining the effect of an anti-platelet agent, clopidogrel, on subsequent events, including bleeding, myocardial infarction, and death after a drug-eluting stent was implanted into a coronary artery. In this population, clopidogrel use may change over time based on a patient's behavior (e.g., nonadherence) and physicians' recommendations (e.g., end of duration of therapy). Consequently, clopidogrel use was treated as a time-varying variable. **Results:** We demonstrate that 1) the sample sizes at three chosen time points are almost identical in the original and weighted datasets; and 2) the covariates between patients on and off clopidogrel were well balanced after stabilized weights were applied to the original samples. **Conclusions:** The stabilized weight-adjusted Kaplan-Meier estimates and modified log-rank and Wilcoxon tests are useful in presenting and comparing survival functions for time-varying treatments in observational studies while adjusting for known confounders.

Keywords: Kaplan Meier estimates, Observational study, Stabilized weights, Stents, Time-varying treatment.

Introduction

Kaplan-Meier estimator of survival functions was developed to account for censoring resulting from incomplete information on outcomes or their timing [1]. Because Kaplan-Meier methods do not control for confounding nor accommodate time-varying treatments, other methods such as parametric survival models and proportional hazards models are often employed [2,3]. Methods using inverse probability of treatment weight (IPTW) have expanded the analytic tools available to researchers to make unbiased comparisons between treatment groups in observational studies [4–6]. Cole et al. [7] used IPTW in a marginal structural left-censored linear model for analyzing the effect of highly active antiretroviral therapy on semiannual repeated assessments of viral load. Cook et al. [8] studied aspirin effects on cardiovascular death in marginal structural discrete-time survival models with time-dependent inverse probability weights. None of those studies used Kaplan-Meier estimates. Xie and Liu [9] developed an adjusted Kaplan-Meier estimator to reduce confounding effects using IPTW for observational studies with time-invariant treatment in which each observation is weighted by its inverse probability of being in a certain group. Their article proposes a weighted log-rank test for comparing survival functions among treatment groups. Sugihara [10] extended this work by proposing IPTW adjusted Kaplan-Meier estimates and a log-rank test that allows comparisons of treatments in settings where there are more than two treatment options.

Several studies have used improved IPTW, namely stabilized weights (SWs), to estimate effects of time-varying treatments in marginal structure models, which involve a range of statistical models, including Cox proportional hazards models for survival data, linear mixed models for repeated measures, and logistic regression models [11–17]. SWs improve on the nonstabilized IPTW methods by reducing the weights of treated subjects with low propensity scores and untreated subjects with high propensity scores. In a recent study [18] describing analysis methods for observational studies with time-invariant treatment, we demonstrated that the use of the SWs 1) produced appropriate estimation of the variance of main effect and maintained an appropriate type I error rate; and 2) yielded appropriate confidence intervals of relative risks in Poisson regression analyses. That study, however, did not explore the influence of SWs when treatments are time-varying.

Time-varying treatments are common. We were motivated to further explore the use of SWs by an observational study

* *Address correspondence to:* Stanley Xu, Kaiser Permanente Colorado - Institute for Health Research, 10065 East Harvard Avenue, Denver, CO 80111 USA.

E-mail: stan.xu@kp.org.

1098-3015/$36.00 – see front matter Copyright © 2012, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

doi:10.1016/j.jval.2011.07.010

where we examined the effects of an anti-platelet medication, clopidogrel, on subsequent risk of bleeding, myocardial infarction (MI), and death after a drug-eluting stent (DES) was implanted into a coronary artery [19]. In that study patients were followed after hospital discharge for up to 18 months. Clopidogrel use was not randomly assigned to patients in that study. Physicians often consider aspects of a patient's clinical condition, such as bleeding history and risk factors for MI and death, when deciding whether or not to initiate clopidogrel therapy and how long to continue clopidogrel treatment. The mean number of clopidogrel-covered days was 257 ± 159 days based on pharmacy dispensing data. Almost half (49.1%) of patients received clopidogrel for longer than 6 months. Because of the reasons above, standard Kaplan-Meier method cannot be applied to assess the risk and benefit of clopidogrel use after DES implantation in this observational study setting. Although a relative-risk-like measure such as the hazard ratio is commonly used to evaluate the effect of treatment, caution about interpreting the hazard ratio need to be taken, especially when it is not consistent over time [20]. Kaplan-Meier estimates and the Kaplan-Meier curves are useful and informative because they reveal the details of changes in survival function over a follow-up period.

In this research we developed a separate probabilistic model of being on clopidogrel for every day of follow-up. Then the corresponding time-dependent SWs were calculated and used to weight both the number of events and the number of subjects at risk during every day of follow-up. We then used this information to develop SW-adjusted Kaplan-Meier estimates and modified log-rank and Wilcoxon tests useful for estimating and comparing survival functions of time-varying treatments in observational studies. Adjusting for covariates when treatment is time-varying in observational studies using SWs, the resulting Kaplan-Meier curves represent survival function estimates for those never treated versus those always treated.

Statistical Methods

Kaplan-Meier estimates when treatment is time-invariant

Let z_i be an indicator of binary treatment with 1 for treated and 0 for untreated for subject i. Suppose that there are n_j subjects at risk at time j, with n_{1j} subjects who received the treatment and n_{0j} subjects who did not, $n_j = n_{0j} + n_{1j}$. In group z, there are d_{zj} events of interest out of n_{zj} subjects. The Kaplan-Meier estimates of survival function for group z at time j is given by $S_{zj} = \prod_{k=1}^{j}(1 - h_{zk})$, where $h_{zj} = d_{zj}/n_{zj}$ is the hazard function at time j [1]. The Kaplan-Meier estimates can be plotted to display the survival function over time for each group. Two common nonparametric statistical tests have been widely used for comparing two groups of survival data: the log-rank test and the Wilcoxon test [21–24].

Crude Kaplan-Meier estimates when treatment is time-varying

Because the treatment is time-varying, we add subscribe j to the indicator of binary treatment, z_{ij}, with 1 for treated and 0 for untreated for subject i at time j. Notations remain the same for the number of subjects at risk at time j (n_{zj}), the number of events at time j (d_{zj}), the hazard function (h_{zj}), and the survival function (S_{zj}) except that subjects may switch their treatment status between treated and untreated at any time during the follow-up. Let y_{zij} be the outcome variable with 1 for event and 0 for no event, and let I_{zij} be an indicator that subject i is at risk in group z at time j. When treatment is time-varying, we propose to calculate the number of events and the number of subjects at risk as $d^c_{zj} = \sum_{i=1}^{n_{zj}} y_{zij} I_{zij}$ and $n^c_{zj} = \sum_{i=1}^{n_{zj}} I_{zij}$, respectively. For each group, the crude Kaplan-Meier estimates for the hazard function for group z at time j can be calculated as

$$h^c_{zj} = \frac{d^c_{zj}}{n^c_{zj}}$$

The corresponding survival probability at time j, given survival to time j-1, is $1 - h^c_{zj}$. The crude Kaplan-Meier estimates of the survival function at time j is

$$S^c_{zj} = \prod_{k=1}^{j}\left(1 - h^c_{zk}\right) \quad (1)$$

SW-adjusted Kaplan-Meier estimates when treatment is time-varying

We first calculate the time-dependent SW for each subject. Let X_{ij} be a row vector of covariates for the probability of treatment and outcome, π_{ij} be the propensity score. For those treated, the propensity score π_{ij} is the probability of treatment given the observed covariates X_{ij}, prob (z_{ij}=1|X_{ij}). For those not treated, it is 1- prob (z_{ij}=1|X_{ij}) [4–6]. The probability of treatment can be estimated with a logistic regression model: $prob(z_{ij} = 1|X_{ij}) = \frac{\exp(X_{ij}\beta_j)}{1+\exp(X_{ij}\beta_j)}$ where β_j is a column vector of parameters to be estimated at time j from data. For the same covariate (fixed or time-varying), estimated values of parameters may vary over time because the status of clopidogrel use changes. At time j for subject i, if z_{ij}=1, then the stabilized weight $W_{ij} = \frac{P_j}{\pi_{ij}}$, and if z_{ij}=0 then $W_{ij} = \frac{1-P_j}{1-\pi_{ij}}$ where p_j is the probability of treatment at time j without considering covariates [16,17]. Theoretically, and by simulation, it has been shown that the use of the SWs in analyzing observational data has appropriate type I error rates when the treatment is time-invariant [18].

To adjust for known covariates, we propose to apply SW to obtain the number of events and the number of subjects at risk as $d^w_{zj} = \sum_{i=1}^{n_{zj}} w_{zij} y_{zij} I_{zij}$ and $n^w_{zj} = \sum_{i=1}^{n_{zj}} w_{zij} I_{zij}$, respectively. For each group, the SW-adjusted Kaplan-Meier estimates of the survival function for group z at time j can be calculated as

$$S^w_{zj} = \prod_{k=1}^{j}(1 - h^w_{zk}) \quad (2)$$

where $h^w_{zj} = \frac{d^w_{zj}}{n^w_{zj}}$ is the corresponding hazard function.

Modified log-rank and Wilcoxon tests for comparing the SW-adjusted survival data of two groups are detailed in the Appendix in Supplemental Materials at: doi:10.1016/j.jval.2011.07.010.

Clopidogrel Use Example

We applied the SW-adjusted Kaplan-Meier estimates, the modified log-rank test and the modified Wilcoxon test to examine the benefit and risk of clopidogrel use for patients who received a DES. Three endpoints, bleeding, MI, and death, are considered in this observational study. The time origin for the survival analysis is day 1 after hospital discharge following DES surgery. In this example, the time-dependent treatment was a dichotomous variable

Table 1 – Sample sizes in the original and weighted data sets.

Days after hospital discharge	Original data set	Weighted data set
1	6447	6418
180	5422	5422
360	4447	4404

with 1 or 0 indicating the use of clopidogrel on that day (i.e., no lag or cumulative functions were used), implying an acute effect of clopidogrel on all endpoints. The following steps were taken to analyze the data. Step 1) fit separate logistic regression models for every day of the follow-up period. At this step the dichotomous variable indicating the use of clopidogrel on that day is the dependent variable; Step 2) calculate the stabilized weights for every day of follow-up period for each individual as defined in the Statistical Methods section; Step 3) calculate the crude and SW-adjusted Kaplan-Meier estimates as in equations (1) and (2); Step 4) obtain the *P* values from the modified log-rank test and Wilcoxon test (see Appendix in Supplemental Materials at: doi:10.1016/j.jval.2011.07.010). We also compared the sample sizes between the original and the weighted data sets. In addition we evaluated the covariates balance between those on and off clopidogrel at every day of the follow-up period. We present the details on covariate balance for three arbitrary time points, 1, 180, and 365 days after hospital discharge.

SWs

Because a patient could discontinue and then restart clopidogrel therapy during the follow-up period, we modeled clopidogrel use as a time-varying dependent variable in creating SW. We fit separate lo-

Table 2 – Comparison of covariates at 1 day after hospital discharge.

Variables	Clopidogrel use in original data			Clopidogrel use in weighted data		
	Off	On	*P*	Off	On	*P*
Age group (%)			<0.0001			0.21
<65	43.0	52.5		48.1	51.0	
≤65 – <75	29.0	27.1		29.9	27.3	
≥75	28.0	20.4		22.0	21.7	
Male (%)	66.2	71.3	0.002	70.9	70.3	0.73
BMI ≥33 (%)	22.9	18.8	0.004	20.8	19.3	0.31
Prior MI (%)	38.0	24.6	<0.0001	28.7	26.9	0.29
Previous valve surgery (%)	1.29	1.22	0.85	0.97	1.23	0.54
Diabetes (%)	40.3	29.1	<0.0001	32.7	30.7	0.26
Renal insufficiency (%)						
1 renal failure and on dialysis	2.70	1.73	0.03	1.53	1.99	0.66
2 renal failure without dialysis	4.11	3.00		3.24	3.34	
3 no renal failure	93.19	95.26		95.2	94.7	
Cerebrovascular disease (%)	14.0	7.4	<0.0001	9.5	8.5	0.38
Peripheral vascular disease	16.43	9.79	<0.0001	11.3	10.7	0.65
Chronic lung disease (%)	16.9	13.9	0.02	13.4	14.4	0.44
Hypertension (%)	81.2	75.2	<0.0001	76.4	76.3	0.91
Tobacco use (%)			0.050			0.91
1 current use	15.6	18.0		17.5	17.6	
2 former use	48.1	43.9		45.1	44.3	
3 never used	36.3	38.1		37.4	38.1	
Hypercholesterolemia (%)	73.5	81.1	<0.0001	79.7	80.4	0.63
Previous revascularization (%)	48.0	29.9	<0.0001	34.6	32.8	0.30
Previous congestive heart failure (%)	80.2	71.7	<0.0001	75.2	72.6	0.12
Cardiogenic shock (%)	3.4	1.4	<0.0001	2.2	1.6	0.29
Use of glycoprotein IIb-IIIa inhibitor (%)			<0.0001			0.98
0 no use	50.2	35.3		38.3	37.9	
1 contraindicated	0.6	0.2		0.3	0.3	
2 yes, used	49.2	64.5		61.4	61.8	
Left ventricular function (%)	59.6	44.3	<0.0001	49.3	46.1	0.08
Discharge location (%)			<0.0001			0.24
1	66.5	71.7		68.7	71.3	
2	23.5	17.0		18.7	17.7	
3	10.0	11.3		12.6	11.0	
Periprocedural MI (%)	0.94	0.95	0.98	1.23	1.04	0.63
No. of stents (%)			0.01			0.46
1	37.4	42.0		38.7	41.6	
2	32.3	29.0		30.3	29.3	
3	12.9	14.3		15.0	14.1	
4+	17.4	14.7		16.0	15.0	
Any lesion complication (%)	3.8	4.5	0.34	4.9	4.5	0.62
1 mm < longest stent size < 21 mm (%)	87.2	92.2	<0.0001	89.9	91.7	0.09
Post-procedure TIMI flow (%)	98.2	97.8	0.43	96.4	97.8	0.006
Off label status (%)	74.8	67.5	<0.0001	72.7	68.0	0.008
Length of stay (mean±SD)	2.4 (3.6)	2.1 (3.2)	0.02	2.3 (3.2)	2.1 (3.4)	0.16
No. of lesions (mean±SD)	1.7 (0.9)	1.7 (0.9)	0.86	1.73 (0.89)	1.66 (0.87)	0.03

BMI, body mass index; MI, myocardial infarction; TIMI, thrombolysis in myocardial infarction.

gistic regression models for every day of follow-up until loss to follow-up or death, while adjusting for a set of known covariates including age, sex, body mass index, prior MI, previous valve surgery, diabetes, renal insufficiency, cerebrovascular disease, peripheral vascular disease, chronic lung disease, hypertension, current tobacco use, previous revascularization, previous congestive heart failure, cardiogenic shock, use of a glycoprotein IIb-IIIa inhibitor upon hospital admission, left ventricular function assessment, discharge location, periprocedural MI, number of stents implanted, any lesion complication, an indicator for the longest stent being between 1 mm and 21 mm, post-procedure thrombolysis in myocardial infarction flow, off-label stent status, length of stay, number of lesions, and an indicator for prior bleeding. Among these variables, indicators for prior bleeding and MI were updated for each successive logistic regression model; other variables are measured at baseline. Off-label stent status is an indicator for off-label use of drug-eluting stents that occurred when the stent was implanted outside of Food and Drug Administration-approved clinical scenarios. Additional details on the variables for this analysis are available at http://www.ncdr.com/WebNCDR/NCDRDocuments/datadictdefsonlyv30.pdf. There were 6447 patients with no missing data for covariates included in the analyses. SWs were then calculated according to the formulas in the Statistical Methods section for each individual on each day of the follow-up period until either the subject was censored or died.

Table 3 – Comparison of covariates at 180 days after hospital discharge.

Variables	Clopidogrel use in original data			Clopidogrel use in weighted data		
	Off	On	P	Off	On	P
Age group (%)			<0.0001			0.50
<65	45.0	53.4		49.8	51.5	
≤65 – < 75	30.4	26.7		28.8	27.5	
≥75	24.6	19.8		21.4	20.9	
Male (%)	68.0	71.2	0.022	71.2	70.5	0.64
BMI ≥33 (%)	17.3	20.4	0.01	19.2	19.6	0.77
Prior MI (%)	26.2	26.2	1.00	25.4	26.6	0.41
Previous valve surgery (%)	1.5	1.1	0.25	1.2	1.2	0.99
Diabetes (%)	29.9	30.0	0.97	30.0	30.2	0.90
Renal insufficiency (%)			0.06			0.34
1 renal failure and on dialysis	2.0	1.5		2.1	1.6	
2 renal failure without dialysis	3.9	2.8		3.6	3.0	
3 no renal failure	94.1	95.7		94.4	95.3	
Cerebrovascular disease (%)	9.2	7.7	0.081	8.3	8.2	0.97
Peripheral vascular disease	11.6	9.7	0.04	10.4	10.2	0.87
Chronic lung disease (%)	16.3	13.3	0.005	14.9	13.9	0.34
Hypertension (%)	75.9	75.8	0.90	75.1	75.8	0.60
Tobacco use (%)			0.83			0.91
1 current use	17.8	17.9		18.1	18.3	
2 former use	45.0	44.1		43.9	44.3	
3 never used	37.2	38.0		38.0	37.4	
Hypercholesterolemia (%)	80.7	79.7	0.43	80.3	80.0	0.83
Previous revascularization (%)	30.8	32.7	0.202	31.8	32.5	0.67
previous congestive heart failure (%)	53.8	77.0	<0.001	71.0	71.2	0.885
Cardiogenic shock (%)	0.9	1.7	0.04	1.7	1.6	0.65
Use of glycoprotein IIb-IIIa inhibitor (%)			0.008			0.94
0 no use	33.1	37.9		37.1	36.6	
1 contraindicated	0.5	0.2		0.3	0.3	
2 yes, used	66.4	61.9		62.6	63.1	
Left ventricular function (%)	37.3	49.1	<0.0001	44.4	46.3	0.23
Discharge location (%)			0.71			0.81
1	70.8	70.6		70.6	70.6	
2	18.7	18.2		19.0	18.4	
3	10.5	11.3		10.4	11.0	
Periprocedural myocardial infarction (%)	0.9	0.7	0.57	0.8	0.7	0.81
No. of stents (%)			<0.0001			0.71
1	48.8	40.0		41.6	42.3	
2	28.3	29.7		30.8	29.2	
3	12.6	14.1		13.2	13.8	
4+	10.2	16.3		14.3	14.7	
Any lesion complication (%)	4.0	4.2	0.69	4.2	4.2	0.96
1 mm < longest stent size < 21 mm (%)	95.3	90.7	<0.0001	91.9	91.7	0.78
Post-procedure TIMI flow (%)	98.4	97.7	0.12	98.1	97.8	0.52
Off label status (%)	59.4	70.2	<0.0001	67.5	67.4	0.94
Length of stay (mean±SD)	2.2 (4.4)	2.1 (2.2)	0.45	2.1 (4.3)	2.0 (2.1)	0.43
No. of lesions (mean±SD)	1.6 (0.8)	1.7 (0.9)	<0.0001	1.6 (0.9)	1.6 (0.9)	0.62

BMI, body mass index; MI, myocardial infarction; TIMI, thrombolysis in myocardial infarction.

Table 4 – Comparison of covariates at 360 days after hospital discharge.

Variables	Clopidogrel use in original data			Clopidogrel use in weighted data		
	Off	On	P	Off	On	P
Age group (%)	47.9	54.8	<0.0001	50.6	52.3	0.37
<65						
≤65 – < 75	30.2	26.2		28.9	27.0	
≥75	21.8	19.0		20.5	20.7	
Male (%)	69.8	71.0	0.42	70.3	71.7	0.32
BMI ≥33 (%)	17.3	22.7	<0.0001	19.0	20.3	0.27
Prior MI (%)	24.9	25.8	0.51	25.5	26.8	0.3
Previous valve surgery (%)	1.0	1.0	0.94	1.1	1.1	0.86
Diabetes %	27.6	31.1	0.01	28.6	30.9	0.08
Renal insufficiency (%)			0.003			0.30
1 renal failure and on dialysis	1.0	2.0		1.3	1.9	
2 renal failure without dialysis	3.5	2.3		3.1	3.2	
3 no renal failure	95.5	95.7		95.5	94.9	
Cerebrovascular disease (%)	7.6	8.1	0.56	8.0	8.2	0.72
Peripheral vascular disease	9.6	9.2	0.62	9.5	9.9	0.63
Chronic lung disease (%)	14.5	12.1	0.02	14.1	12.3	0.2
Hypertension (%)	76.2	75.4	0.55	75.7	75.3	0.73
Tobacco use (%)			0.02			0.51
1 current use	18.0	18.0		17.7	19.0	
2 former use	45.3	41.4		43.4	42.6	
3 never used	36.7	40.7		38.9	38.4	
Hypercholesterolemia (%)	82.7	76.4	<0.0001	81.2	79.6	0.17
Previous revascularization (%)	29.4	33.2	0.006	31.0	32.4	0.32
previous congestive heart failure (%)	55.5	82.0	<0.0001	66.6	70.6	0.003
Cardiogenic shock (%)	1.1	2.0	0.02	1.4	1.6	0.67
Use of glycoprotein IIb-IIIa inhibitor (%)			0.03			0.68
0 no use	34.1	37.3		36.7	35.5	
1 contraindicated	0.4	0.2		0.3	0.3	
2 yes, used	65.5	62.5		63.0	64.3	
Left ventricular function (%)	37.7	55.1	<0.0001	45.1	47.1	0.20
Discharge location (%)			0.001			0.49
1	71.0	65.8		70.0	68.3	
2	18.9	21.0		19.0	20.2	
3	10.1	13.2		11.0	11.5	
Periprocedural myocardial infarction (%)	0.9	0.6	0.31	0.9	0.7	0.62
No. of stents (%)			<0.0001			0.32
1	46.2	37.6		42.9	41.4	
2	28.6	29.7		29.7	28.6	
3	13.4	14.5		13.4	14.4	
4+	11.8	18.2		14.1	15.6	
Any lesion complication (%)	3.9	4.7	0.18	4.1	4.5	0.43
1 mm < longest stent size < 21 mm (%)	95.2	89.9	<0.0001	93.2	92.1	0.14
Post-procedure TIMI flow (%)	98.3	97.2	0.01	97.9	97.7	0.57
Off label status (%)	61.3	72.4	<0.0001	65.4	67.6	0.14
Length of stay (mean±SD)	2.0 (3.5)	2.1 (2.3)	0.22	2.1 (4.1)	2.1 (2.1)	0.97
No. of lesions (mean±SD)	1.6 (0.9)	1.7 (0.9)	0.20	1.7 (0.9)	1.7 (0.9)	0.17

BMI, body mass index; MI, myocardial infarction; TIMI, thyrombolysis in myocardial infarction.

Regardless of endpoints, we used the daily SWs to examine the sample sizes (Table 1) and covariates balance at three arbitrarily chosen time points: day 1, 180, and 360 out of 540 days after hospital discharge (Tables 2–4). Because SW-adjusted Kaplan-Meier estimates in equation (2) and modified log-rank and Wilcoxon tests (see Appendix in Supplemental Materials at: doi:10.1016/j.jval.2011.07.010) for an endpoint depend on only the SWs on days with the specific endpoint event, different SWs are applied for each distinct endpoint and are re-estimated each time an endpoint occurs.

Sample sizes in the weighted data sets

Table 1 shows the sample sizes in the original and weighted data sets at day 1, 180, and 360 out of 540 days after hospital discharge. These results demonstrate that the use of stabilized weights preserves the original sample size and thus the use of SWs would not be expected to inflate type I error rates in relevant statistical hypothesis tests including the modified log-rank and Wilcoxon tests for comparing survival data of two groups.

Covariate balance between those on and off clopidogrel use

SW weightings improved covariate balance in analyzing the clopidogrel use data. During the 540 days of follow-up, in the original cohort the mean number of unbalanced covariates (statistically significant based on $P < 0.05$) was 13.5 with a minimum of eight and a maximum of 20 unbalanced covariates out of 27, whereas in the SW weighted cohort, the mean number of unbalanced covari-

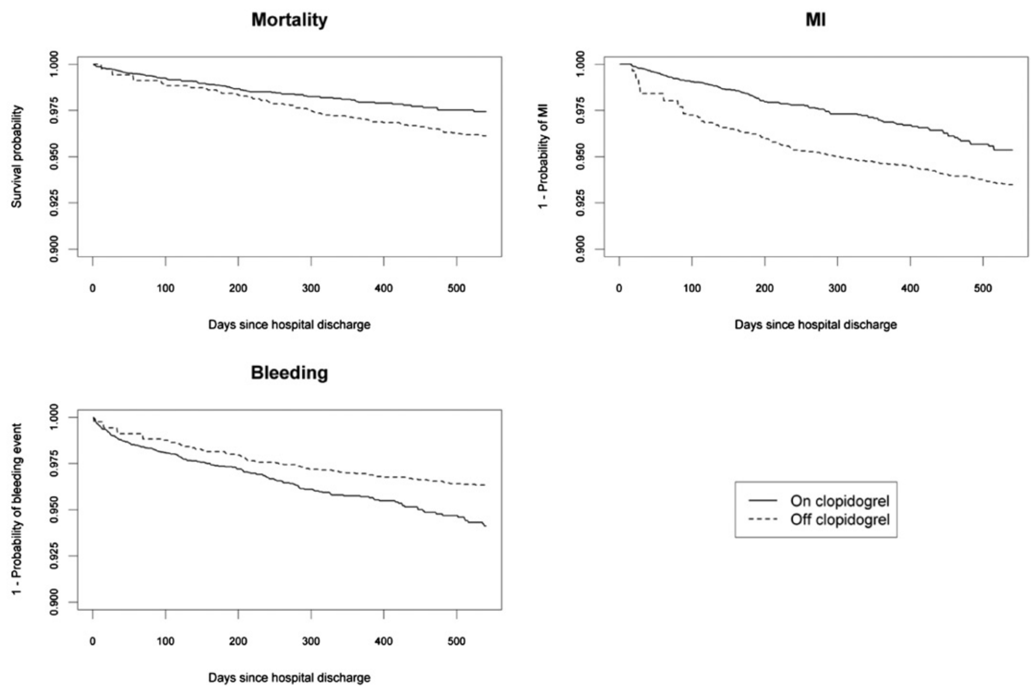

Fig. 1 – Crude Kaplan-Meier curves for bleeding, myocardial infarction (MI), and death.

ates was 1.2 with a minimum of zero and a maximum of five unbalanced covariates out of 27. A mean of 1.2 (out of 27) unbalanced covariates is consistent with chance, based on a 0.05 type I error rate. For example, with SWs applied at day 1 after hospital discharge, 21 unbalanced covariates in the original cohort became balanced; among the remaining six covariates, three balanced covariates in the original cohort remained balanced, two balanced covariates in the original cohort became unbalanced, and one unbalanced covariate in the original cohort remained unbalanced (Table 2). With SWs applied at day 180 after hospital discharge, 13 unbalanced covariates in the original cohort became balanced, the remaining 14 balanced covariates in the original cohort remained balanced (Table 3). With SWs applied at day 360 after hospital discharge, 16 unbalanced covariates in the original cohort became balanced; among the remaining 11 covariates, 10 balanced covariates in the original cohort remained balanced, one unbalanced covariates in the original cohort remained unbalanced (Table 4).

Kaplan-Meier estimates of survival functions

Crude and SW-adjusted Kaplan-Meier estimates of survival functions for bleeding, MI, and death are displayed in Figures 1 and 2. The corresponding *P* values based on the crude and modified log-rank and Wilcoxon tests are shown in Table 5. Although both the crude and SW-adjusted methods showed that clopidogrel use significantly increased the risk of bleeding and decreased the risk of death during the 540 days of follow-up after hospital discharge, the latter method provided smaller *P* values. Results using the crude log-rank and Wilcoxon tests suggested there was no statistically significant reduction in MI incidence associated with clopidogrel use. The SW-adjusted log-rank and Wilcoxon tests, however, showed a statistically significant reduction of MI incidence associated with clopidogrel use.

We also fit Cox regression models with time-dependent clopidogrel treatment [25] adjusting for the same covariates used in creating SWs. Clopidogrel use significantly increased the risk of bleeding (hazard ratio 1.52; $P = 0.010$) and decreased mortality (hazard ratio 0.67; $P = 0.018$). Although not statistically significant, it also decreased the risk of MI (hazard ratio 0.78; $P = 0.080$).

Discussion

This research extends prior work on IPTW adjusted Kaplan-Meier estimates [9] to include time-varying treatments. Equations for modified log-rank and Wilcoxon tests of survival functions were also detailed and explored using a specific study example. Our study demonstrated that the sample sizes at three time points were almost identical in the original and weighted datasets, suggesting that the use of SWs would not be expected to inflate type I error rates. The balance in the covariates between patients on and off clopidogrel was largely achieved after SWs were applied to the original sample. At each event time *j*, for the entire population of patients who survived to time *j-1*, the survival probabilities for treated and not treated are calculated after covariates are balanced by applying SWs. Because re-estimating probabilities of treatment removes differences between persons on and off treatment and prior history of treatment does not influence the calculation, the switch between on and off clopidogrel at time *j* from time *j-1* would not change the interpretation of the survival probabilities for the two groups at time j when compared to those from survival analysis with time-invariant treatment. The resulting cumulative survival probabilities as displayed in Kaplan-Meier curves represent the estimates for groups that are always treated or never treated. Thus, we believe that SW-adjusted Kaplan-Meier estimates and the modified log-rank and Wilcoxon tests presented here are proper for examining the risk and benefit of clopidogrel use after DES implantation with time-varying treatment adjusting for covariates.

In creating SWs, logistic regression models were fit for every day of the follow-up period so that the covariate balance could be evaluated and all three outcomes of clopidogrel use data could be analyzed with the same data steps. Fitting logistic regression models for every day of follow-up periods could be too time-consuming for studies that have a long follow-up period. SW-adjusted

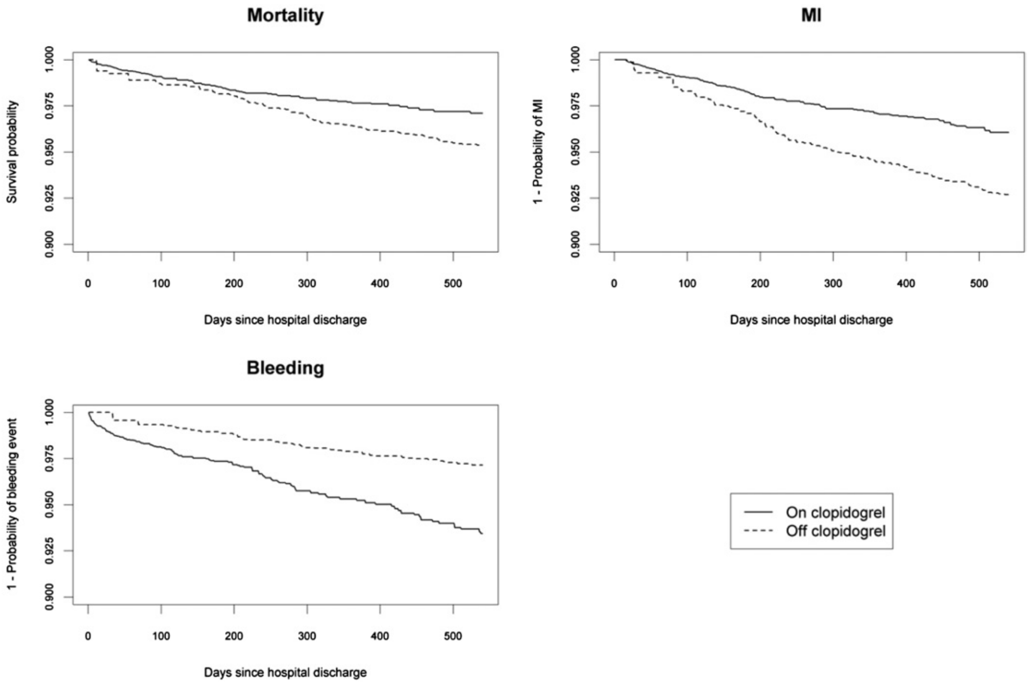

Fig. 2 – Stabilized weight adjusted Kaplan-Meier curves for bleeding, myocardial infarction (MI), and death.

Kaplan-Meier estimates in equation (2) and modified log-rank and Wilcoxon tests (Appendix in Supplemental Materials at: doi: 10.1016/j.jval.2011.07.010) do not depend on the SWs on days without events. In the other words, only the SWs on event days need be calculated, thus reducing the number of logistic regression models necessary to create SWs.

The log-rank test and the Wilcoxon test of survival functions can produce disparate estimates of significance because the two tests weight time periods of follow-up differently [3]. The log-rank test weights the entire follow-up period equally, whereas the Wilcoxon test weights by the sample size at each time point and therefore weights earlier times more heavily. In the example described in this article, small differences in significance levels were noted for the two tests although the conclusions remained consistent. When disparate results do occur, the differences can provide researchers with insights as to where the survival functions differ [3].

Although the clinical conclusions from the Cox regression models adjusting for the same covariates with time-dependent clopidogrel treatment remain the same as from SW-adjusted Kaplan-Meier approach, the *P* values may not be the same. For example, for the MI outcome, the Cox model yielded a marginally significant *P* value of 0.08 whereas the SW-adjusted log-rank test and Wilcoxon test produced *P* values less than 0.0001. In the Cox model, covariates are included as predictors of the outcome along with the indicator for clopidogrel use, whereas in the SW-adjusted Kaplan-Meier approach the covariates are treated as confounders that are associated with both the outcome and the clopidogrel use status. Also the Cox model requires a proportional hazard assumption [3], but the SW-adjusted Kaplan-Meier approach does not. In the presence of strong confounding or violation of the proportional hazard assumption, results from the Cox model may not be valid. For example, the Kaplan-Meier curves for MI in Figure 1 clearly suggest nonproportional hazards between the two groups, with the Off Clopidogrel group dropping rapidly within the first 50 days, and paralleling the On Clopidogrel group thereafter.

Examination of covariate balance in this study was limited. We provide a summary of covariate balance over the entire follow-up period, and show covariate differences at three arbitrary time points using p-values to demarcate potential imbalance. *P* values are affected by sample size, do not adequately focus on actual difference magnitudes, and have other well-documented shortcomings [26]. Time-varying covariates can result in the need for large number logistic regression iterations, as in the example study where 540 treatment models were fit. Examination of the actual magnitude of difference across all covariates for each individual logistic regression alone in such instances is impractical. For the three time periods shown here, some of the covariate balance differences were likely related to changes in the proportion of persons treated with clopidogrel on that day. At day 1 after hospital discharge, only 15% of patients were not on clopidogrel, whereas 25% and 52% were not on clopidogrel at 180 days and 360 days after hospital discharge, respectively. Clearly, assessing and confirming adequate covariate balance in IPTW time-varying models is challenging and needs further study. An additional limitation of this research is the single study application. Further work with simulations and contrasts to other methods and other study applications would help elucidate the advantages and disadvantages of this approach.

Table 5 – *P* values for bleeding, myocardial infarction (MI), and death based on crude and stabilized weights (SW) adjusted log-rank and Wilcoxon tests.

	Bleeding		MI		Death	
	Crude	SW	Crude	SW	Crude	SW
Log-rank	0.0056	0.0002	0.136	<0.0001	0.0079	0.0020
Wilcoxon	0.0150	0.0003	0.052	<0.0001	0.0106	0.0033

Conclusions

This article examined the use of SWs in Kaplan-Meier estimates with time-varying treatments in observational studies. Related methods include marginal structural logistic models that, for example, have been used to explore the effect of iron supplementation on anemia in pregnancy [27] and marginal structural Cox models that have investigated aspirin effects on cardiovascular death [8]. The Kaplan-Meier method explored here expands on the available analytic tools and has the advantage of aligning more directly to the adjusted survival graphing methods detailed by Cole and Hernán [28]. Analyses of observational studies will continue to be improved by further detailed exploration of these related methods and further exploration of their use.

Source of financial support: This project was supported by Strategic Initiatives Funds of Kaiser Permanente Colorado, supported by NIH/NCRR Colorado CTSI grant No. TL1 RR025780, and funded under Contract No. 290-05-0033 from the Agency for Healthcare Research and Quality, US Department of Health and Human Services as part of the Developing Evidence to Inform Decisions about Effectiveness 16 program.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at doi:10.1016/j.jval.2011.07.010, or if hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53,457–81.
[2] Cox DR. Regression models and life tables. J R Stat Soc Series B 1972; 34:187–220.
[3] Hosmer DW, Lemeshow S. Applied Survival Analysis. Regression Modeling of Time to Event Data. New York: John Wiley & Sons, 1999.
[4] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
[5] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79:516–24.
[6] D'Agostino RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–81.
[7] Cole SR, Hernàn MA, Anastos K, et al. Determining the effect of highly active antiretroviral therapy on changes in human immunodeficiency virus type 1 RNA viral load using a marginal structural left-censored mean model. Am J Epidemiol 2007;166:219–27.
[8] Cook NR, Cole SR, Hennekens CH. Use of a marginal structural model to determine the effect of aspirin on cardiovascular mortality in the Physicians' Health Study. Am J Epidemiol 2002;155:1045–53.
[9] Xie J and Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Stat Med 2005;24:3089–110.
[10] Sugihara, M. Survival analysis using inverse probability of treatment weighted methods based on the generalized propensity score. Pharmaceut Stat 2010;9:21–34.
[11] Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiol 2000;11:561–70.
[12] Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of non-randomized treatments. J Am Stat Assoc 2001;96:440–8.
[13] Hernán MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. Stat Med 2002;21:1689–709.
[14] Wang C, Vlahov D, Galai N, et al. The effect of HIV infection on overdose mortality. AIDS 2005;19:935–42.
[15] Sterne JA, Hernán MA, Ledergerber B, et al. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. Lancet 2005;366:378–84.
[16] Robins JM. Marginal structural models. 1997 Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association, 1998;1–10.
[17] Robins JM, Hernán M, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiol 2000;11:550–60.
[18] Xu S, Ross C, Raebel M, et al. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. Value Health 2010;13:273–7.
[19] Tsai TT, Ho PM, Xu S, et al. Increased risk of bleeding in patients on clopidogrel therapy after drug-eluting stents implantation: insights from the HMO Research Network-Stent Registry (HMORN-Stent). Circ Cardiovasc Interv 2010;3:230–5.
[20] Hernán MA. Hazards of hazard ratios. Epidemiol 2010; 21:13–5.
[21] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 1966;50: 163–70.
[22] Peto R, Peto J. Asymptotically efficient rank invariant test procedures. J R Stat Soc Series A 1972;135:185–207.
[23] Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika 1965;52:203–23.
[24] Breslow N. A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika 1970;57:579–94.
[25] Andersen P, Gill R. Cox's regression model for counting processes, a large sample study. Ann Stat 1982;10:1100–20.
[26] Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3d ed. Philadelphia: Lippincott Williams & Wilkins, 2008.
[27] Bodnar LM, Davidian M, Siega-Riz AM, Tsiatis AA. Marginal structural models for analyzing causal effects of time-dependent treatments: an application to perinatal epidemiology. Am J Epidemiol 2004;159:926–34.
[28] Cole SR, Hernàn MA. Adjusted survival curves with inverse probability weights. Comput Methods Programs Biomed 2004;75:45–9.